Online Adaptive Multiple Appearances Model for Long-Term Tracking

Shuo Tang, Longfei Zhang(✉), Xiangwei Tan, Jiali Yan, and Gangyi Ding

Digital Performance and Simulation Key Laboratory, School of Software,
Beijing Institute of Technology, Beijing 100081, China
{shuo_tang,longfeizhang,dgy}@bit.edu.cn

Abstract. How to build a good appearance descriptor for tracking target is a basic challenge for long-term robust tracking. In recent research, many tracking methods pay much attention to build one online appearance model and updating by employing special visual features and learning methods. However, one appearance model is not enough to describe the appearance of the target with historical information for long-term tracking task. In this paper, we proposed an online adaptive multiple appearances model to improve the performance. Building appearance model sets, based on Dirichlet Process Mixture Model (DPMM), can make different appearance representations of the tracking target grouped dynamically and in an unsupervised way. Despite the DPMM's appealing properties, it characterized by computationally intensive inference procedures which often based on Gibbs samplers. However, Gibbs samplers are not suitable in tracking because of high time cost. We proposed an online Bayesian learning algorithm to reliably and efficiently learn a DPMM from scratch through sequential approximation in a streaming fashion to adapt new tracking targets. Experiments on multiple challenging benchmark public dataset demonstrate the proposed tracking algorithm performs 22 % better against the state-of-the-art.

Keywords: Object tracking · Multiple appearance model · Online Dirichlet process mixture model

1 Introduction

Object tracking plays an important role in numerous vision applications, such as motion analysis, activity recognition, visual surveillance and intelligent user interfaces. However, while much progress has been made in recent years, it is still a challenging problem to track a moving object in a long term in the real-world because of the variations of tracking environment such as view port exchanging, illuminance varying, and etc. For visual tracking problem, an appearance model is used to represent the target object and predicted the likely states of tracking target in future frame [1]. However, using one appearance model is not suitable to describe all the historical appearance information, especially for long term tracking task. So we mainly focused on building multiple appearance models

T. Tan et al. (Eds.): CCPR 2016, Part I, CCIS 662, pp. 501–516, 2016.
DOI: 10.1007/978-981-10-3002-4_42

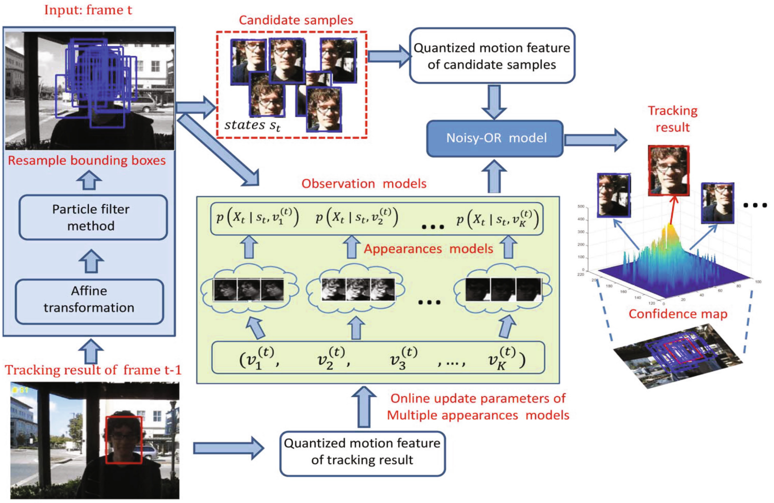

Input: frame t
Candidate samples
states s_t
Quantized motion feature of candidate samples
Noisy-OR model
Tracking result
Resample bounding boxes
Particle filter method
Affine transformation
Observation models
$p(X_t \mid s_t, v_1^{(t)})$ $p(X_t \mid s_t, v_2^{(t)})$ ... $p(X_t \mid s_t, v_K^{(t)})$
Appearances models
...
$(v_1^{(t)}, v_2^{(t)}, v_3^{(t)}, \ldots, v_K^{(t)})$
Confidence map
Tracking result of frame t-1
Online update parameters of Multiple appearances models
Quantized motion feature of tracking result

Fig. 1. The framework of online adaptive multiple appearances model based tracking.

dynamically in an unsupervised way in order to adapt the changing appearance of tracking target (Fig. 1).

In this paper, we used Bayesian non-parametric clustering method to cluster multiple different appearances dynamically. It can cover more aspects of the target appearance to make the proposed algorithm more robust to abrupt appearance changes, and the number of clusters can be inferred from the observation. Among the different probabilistic models, Bayesian non-parametric method has several properties which suit the object tracking application well. In particular, DPMM represents mixture distributions with an unbounded number of components where the complexity of the model adapts to the observed data. This property is important for building multiple appearance models dynamically. In general, the number of appearances is uncertain and varying over time.

However, despite the appealing properties of DPMM, it characterized by computationally intensive inference procedures, which often based on Gibbs samplers [2]. While Gibbs sampling can be an appropriate inference mechanism when execution time is not an issue. It is not applicable in visual tracking, as it needs more faster inference. In [3] a variational inference method which maximizes a lower bound to the true underlying distribution and after each iteration, the obtained parameters define a distribution which approximates the true one in a properly defined way. However, variational inference method is extremely vulnerable to local optima for non-convex unsupervised learning problems, and is frequently yielding poor solutions.

In visual tracking literature, the appearance model based on BNP is not applied as usual as the parametric methods [4]. The main strategy of the BNP tracking methodology is based on three aspects shown as follows: we need to solve (1) how to represent the observation of tracking target by Bayesian non-parametric models, (2) how to create multiple appearance models dynamically without knowledge of cluster numbers and model parameters to adapt in tracking environment variation, (3) how to update multiple appearance models effectively and reliably in tracking process.

Our proposed algorithm is mostly inspired by [5,6] which are online Bayesian learning algorithm to estimate DP mixture models. This method does not require random initialization like Gibbs samplers. Instead, it can reliably and efficiently learn a DPMM from scratch through sequential approximation in a single pass. The algorithm takes data in a streaming fashion, and thus can be easily adapted to new tracking target.

The rest of the paper is organized as follows: Sect. 2 reviews some of the related work. Section 3 reviews Bayesian non-parametric model on which our proposed model based. Section 4 introduces multiple appearances modeling and representation, and proposes related probabilistic distributions which can describe the generation process of tracking features. Section 5 introduces an online sequential Bayesian method to build multiple appearance models. Section 6 presents the framework of the proposed tracking algorithm. Section 7 reports the experimental results. Section 8 makes the conclusion of the paper.

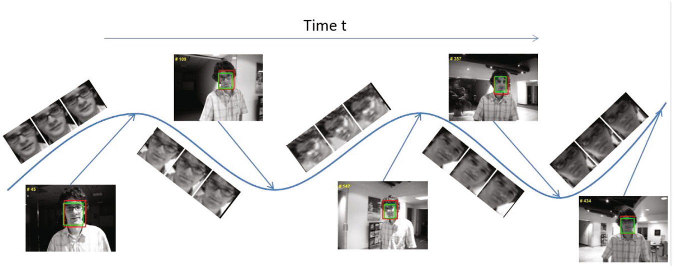


Fig. 2. The appearances when operating the online adaptive multiple appearances model tracking. (Color figure online)

2 Related Work

There is a rich literature in object tracking approatches [7]. As a main component in tracking algorithms, tracking target appearance modeling plays a key role in tracking performance. A good appearance representation should have strong description or discrimination power to distinguish the target from the background. In order to adapt to the appearance variations of the target during tracking, there are many adaptive appearance models have been proposed for object tracking including both generative and discriminative methods.

For generative appearance modeling methods, Jepson et al. [8] learn a Gaussian mixture model via an online expectation maximization algorithm to account for target appearance variations during tracking. Incremental subspace methods have also been used for online object representation [9]. This method uses target observations obtained online to learn a linear subspace for object representation. Since the appearance of a target in a long-time interval may be quite different, these generative models may not describe the appearance variations of the target well.

For discriminative appearance modeling methods, Avidan et al. [10] use online boosting method for tracking. They proposed an ensemble tracking framework to construct a strong classifier to distinguish the target from the background. Babenko et al. [11] use Multiple Instance Learning (MIL) instead of traditional supervised learning to avoid the inaccuracy accumulation problem caused by self-learning. In these methods, tracking is usually treated as a binary classification problem. In order to train and update the classifiers, samples usually needs to be correctly labeled, which may not be available in many real tracking applications.

The most related methods to our model is [12], which proposed the original Adaptive Multiple Appearance Model (AMAM) framework to maintain not only one appearance model as many other tracking methods but appearance model set to describe all historical appearances of the tracking target during a long term tracking task. This method employed DPMM to build multiple appearance models unsupervised to tackle drifting problem, and experiment in several public datasets shows that this tracker has high tracking performance compared with several other state-of-the-art. In order to infer the number of different appearances underlying tracking observations, this tracker resorts to Gibbs sampler [2] for approximate inference and also requires random initialization of components. However, as this sampler needs to maintain the entire configuration, the computational complexity of this tracker is quite high, which limits its applications in real-time scenarios.

Compared with the tracking methods as described above, our proposed method shows three mainly characteristics in dealing with appearance variations of the target. Firstly, our method can cluster multiple different appearances dynamically, and the number of clusters can be inferred from the tracking observation. Secondly, in our method, different kinds of tracking target appearances can be modeled by new model or constructed appearance models. It covered various target appearances, which made the proposed method more robust to abrupt appearance changes. Finally, our method begins with an empty model and progressively refines the models as tracking observation come in, adding new appearance models on the fly when needed.

3 Bayesian Non-parametric Model

The Dirichlet process (DP) introduced in [14], is a popular nonparametric stochastic process that defines a distribution over probability distributions. The DP is parameterized by a base distribution H which has corresponding density $h(\mu)$, and a positive scaling parameter $\alpha > 0$. We denote a DP as follows:

$$G|\{\alpha, H\} \sim DP(\alpha, H), G \triangleq \sum_{k=1}^{\infty} \omega_k \delta_{\phi_k},$$

$$v_k \sim Beta(1, \alpha), \omega_k = v_k \prod_{l=1}^{k-1} (1 - v_l)$$

The DP is most commonly used as a prior distribution on the parameters of a mixture model when the number of mixture components is uncertain. Such a model is called a Dirichlet process mixture model (DPMM) which can be specified as:

$$G \sim DP(\alpha, H), \theta_i | G \sim G, x_i | \theta_i \sim F(\theta_i), i \in \{1, \cdots, N\}$$

Let z_i indicate the subset, or cluster, associated with the i^{th} observation, the DP mixture model can also be modeled by using the Chinese Restaurant Process (CRP) representation [17] of the DP, leading to the followings:

$$p(z_i = k | z_{-i}) \propto \sum_{j \neq i} 1_{z_j = k}, p(z_i = k_{new} | z_{-i}) \propto \alpha_0$$

So, a model equivalent to the DPMM using the CRP can be specified as:

$$z \sim CRP(\alpha), \theta_k | G \sim G, x_i | z_i \sim F(\theta_{z_i})$$

4 Multiple Appearance Modeling and Representation

In this section, our goal is to develop a probabilistic method to cluster multiple different appearances unsupervised, which can cover aspects of the target appearance. In order to do so, we represent motion features (e.g. HOG, color feature etc.) using histograms, and then, quantize motion feature values of tracking observations to 20 or more levels, which is a common practice for similar histogram-based descriptors, such as [13]. Thus, considering N tracking observations $X = \{X_i\}_{i=1}^{N}$, which can be clustered into K clusters or different appearances, and each $X_i = \{x_i\}_{i=1}^{D}$ represents a quantized D dimensional motion feature, and x_i is the corresponding histogram quantized bin counts, which is a quantized integer. With the new tracking observation arrival, the number of clusters became to be varies. Given the cluster assignment for i^{th} each observation X_i, its likelihood for that cluster is $F(X_i|\theta_k)$, while the $\theta_{1:K}$ are drawn from the base distribution of DPMM.

4.1 Exponential Family and Sufficient Statistics

In order to describe motion feature X which is the collection of small integers of histograms and the histogram bin counts, we adopt component distributions of which are members of the exponential family distributions. The base measure

Algorithm 1. The method of building multiple appearance models

```
Input: Given the concentration parameter $\alpha$ of $DPMM$, base measure
parameter $\mu = (\mu_1, \cdots \mu_D)$ of $H(\mu)$ and HOG features $\{X_i\}_{i=1}^N$ of tracking
observations, until frame $N$.
Output: multiple appearance model parameters $\zeta_k, (k = 1 : K)$
Let $K = 1, \rho_1(1) = 1, \omega_1 = \rho_1, \zeta_1 = \mu$
for $i = 2 \quad to \quad N$ do
    Compute the marginal likelihood $f_k(X_i) = p(X_i|\zeta_k)$, for $k = 1 : K$
    Compute $f_k(X_i) = p(X_i|\lambda)$using Eq. (3), for $k = K + 1$
    Compute $\rho_i(k) = \omega_k f_k(X_i) / \sum_l \omega_l f_l(X_i)$ for $k = 1 : K + 1$ , with $\omega_{K+1} = \alpha$
    see, Eq. (9)
    if $\rho_i(K + 1) > \varepsilon$ then
        for $k = 1 : K$ do
            $\omega_k = \omega_k + \rho_i(k)$,
            Update parameters according to posterior and
            $\rho_i(k) : \theta_k = \theta_k + X_i \rho_i(k)$ using Eq. (10)
        $\omega_{K+1} = \rho_i(K + 1)$ , $\zeta_{K+1} = X_i \zeta_i(K + 1)$
        $K = K + 1$
    else
        Re-normalize $\rho_i$ such that $\sum_{k=1}^K \rho_i(k) = 1$
        $\omega_k = \omega_k + \rho_i(k), \zeta_k = \zeta_k + X_i \rho_i(k)$ for $k = 1 : K$
```

of the DPMM will be the conjugate prior, because it has many well-known properties, which can admit efficient inference algorithms. Thus, in this paper, we will consider to describe the distributions as follows:

$$p(X_n|\theta_i) = l(X_n) exp(\theta_i^T X_n - a(\theta_i)),$$

where a is the log-partition function. We take H to be in the corresponding conjugate family:

$$h(\theta|\lambda) = l(\theta) exp(\lambda_1^T \theta - \lambda_2 a(\theta) - a(\lambda)),$$

where the sufficient statistics are given by the vector $(\theta^T, -\alpha(\theta))$, and $\lambda = (\lambda_1^T, \lambda_2)$.

4.2 Model Representation

Specifically, we choose Multinomial distribution $F(X_i|\theta)$, which we denote $Mult(\theta_k; n)\theta_k = (p_1, \cdots, p_D)$, is a discrete distribution over D dimensional non-negative integer vectors $X_i = (x_1, x_2, \cdots x_D)$ where $\sum_{i=1}^D x_i = n$. The probability mass function is given as follows:

$$f(X_i; p_1, \cdots, p_D, n) = \frac{\Gamma(n+1)}{\prod_{i=1}^D \Gamma(x_i + 1)} \prod_{i=1}^D p_i^{x_i} \qquad (1)$$

The cluster prior $H(\theta|\lambda)$ is represented by a Dirichlet distribution which is conjugate to $F(X_i|\theta)$. We denote cluster prior $H(\theta|\lambda)$ as follows, which is a Dirichlet distribution and is conjugate to $F(X_i|\theta)$.

$$h(p_1, \cdots p_D; \lambda_1, \cdots \lambda_D, n) = \frac{1}{B(\lambda_1, \cdots \lambda_D)} \prod_{i=1}^{D} p_i^{\lambda_i - 1}, \tag{2}$$

where the normalizing constant is the multinomial Beta function. Because $H(\lambda)$ is conjugate to $F(\theta)$, then the marginal joint distribution can be obtained by integrating out $(p_1, \cdots, p_D)$ as follows:

$$p(X_i|\lambda_1, \cdots \lambda_D) = \frac{N!}{\prod_{i=1}^{D}(n_i!)} \frac{\Gamma(A)}{\Gamma(N+A)} \prod_{i=1}^{D} \frac{\Gamma(n_i + \lambda_i)}{\Gamma(\lambda_i)} \tag{3}$$

where $A = \Sigma_i \lambda_i$ and $N = \Sigma_i n_i$, and where n_i = number of x_i's with value i.

4.3 Multiple Appearance Modeling

When the number of clusters K is estimated, the multiple appearance model can be built. Considering all of the model parameters, which is comprised of the model parameters $\theta_{1:K}$ and the cluster indicator $z_{1:N}$, the joint distribution of this Bayesian Non-parametric mixture model can be written as in Eq. (4).

$$\begin{gathered} p(\theta_{1:K}, z_{1:N}|X_{1:N}) \propto \\ p(z_{1:N})(\prod_{i=1}^{N} p(X_i|\theta_{z_i})) \prod_{k=1}^{K} p(\theta_k) \end{gathered} \tag{4}$$

Here, $z_i \in \{1 \cdots K\}$ with $i \in \{1 \cdots N\}$ indicates the cluster label of the observation X_i and θ_k are the parameters for the k-th appearance model. The target of our proposed method is to infer the joint posterior distribution $p(\theta_{1:K}, z_{1:N}|X_{1:N})$ unsupervised and dynamical, then we can get the parameters $\theta_{1:K}$ of multiple appearance models.

5 Online Sequential Approximation

In order to infer the joint posterior distribution $p(\theta_{1:K}, z_{1:N}|X_{1:N})$, we can initialize the components randomly, and then resort to Gibbs sampler for approximate inference [12]. However, this method needs to maintain the entire configuration, so the computational complexity of this tracker is rather high, which limits its applications in real-time scenarios.

We improved it by using an online sequential variational approximation method to learn a DPMM from scratch through sequential approximation in a streaming, which is easily adapted to new observation.

By marginalizing out the cluster assignment $z_{1:N}$, we obtain the posterior distribution $p(\theta|X_{1:N})$:

$$p(\theta|X_{1:N}) = \sum_{z_{1:n}|X} p(z_{1:N}|X_{1:N})p(\theta|X_{1:N}, z_{1:N}) \tag{5}$$

In order to compute the distribution above, it requires enumerating all possible partitions $z_{1:N}$, which grows exponentially as n increases. To tackle this difficulty, we resort to variational approximation [6] to choose a tractable distribution to approximate $p(\theta|X)$ as follows:

$$q(\theta|\rho, v) = \sum_{z_{1:n}} \prod_{i=1}^{n} \rho_i(z_i) q_v^{(z)}(\theta|z_{1:n}) \tag{6}$$

We begin our tracker with one appearance model (i.e. $K = 1$) and progressively refine the model as samples come in, adding new appearance models on the fly when needed. Specifically, when we have $\rho = (\rho_1, \rho_2, \cdots, \rho_i)$ and $v^{(i)} = (v_1^{(i)}, v_2^{(i)}, \cdots, v_K^{(i)})$ after processing i frames. To determine X_{i+1}, we can use either of the K existing appearance models or generate a new model θ_{K+1}. Then the posterior distribution of $z_{i+1}, \theta_1, \cdots, \theta_{K+1}$ given $x_1, \cdots, x_{i+1}$ is

$$\begin{aligned} p(z_{i+1}, \theta_{1:K+1}|X_{1:i+1}) \propto \\ p(z_{i+1}, \theta_{1:K+1}|X_{1:i})p(X_{i+1}, z_{i+1}|\theta_{1:K+1}) \end{aligned} \tag{7}$$

Using the tractable distribution $q(\theta|\rho, v)$ to approximate the posterior $p(z_{i+1}, \theta_{1:K+1}|X_{1:i})$, we get the following:

$$p(z_{i+1}, \theta_{1:K+1}|X_{1:i+1}) \propto q(z_{i+1}|\rho_{1:i}, v^{(i)})p(X_{i+1}|z_{i+1}\theta_{1:K+1}) \tag{8}$$

Then, for our model, the optimal setting of q_{i+1} and $v^{(i+1)}$ minimizes the Kullback-Leibler divergence between $q(z_{i+1}, \theta_{1:K+1}|\rho_{1:i+1}, v^{(i+1)})$ and the approximate posterior in Eq. (8) are given as follows:

$$\rho_{i+1} \propto \begin{cases} \omega_k^{(i)} \int F(X_{i+1}|\theta)v_k^{(i)}(d\theta) & (k \leq K) \\ \alpha \int F(X_{i+1}|\theta)h(d\theta) & (k = K+1), \end{cases} \tag{9}$$

with $\omega_k^{(i)} = \sum_{j=1}^{i} \rho_j(k)$, and

$$v_k^{(i+1)}(\theta) \propto \begin{cases} h(\theta) \prod_{j=1}^{i+1} F(X_j|\theta)\rho_j(k) & (k \leq K) \\ h(\theta)F(X_{i+1}|\theta)\rho_{i+1}(k) & (k = K+1) \end{cases} \tag{10}$$

Algorithm 1 illustrates the basic flow of our algorithm. More details can be found in [5]. The implementation of this algorithm is under the circumstance where H and F are exponential family distributions that form a conjugate pair. In such cases, base measure h and posterior measures v_k can be represented by natural parameter denoted by λ and ζ_k.

Algorithm 2. A Summary of the proposed tracking method

```
(1)$L_t(X) \in R_2$ denotes the location of sample $X$ at the $t$-th frame. We have the object location $L_t(X)$ where we assume the corresponding sample is $X_t$ representing the quantized HOG feature.
(2)We apply the affine transformation to $L_t(X)$ with six affine parameters to product candidate samples $s_t$.
(3)For each candidate samples $s_t$, we extract quantized HOG feature$X_t$, then use NOR model of Eq. 13 and each of the multiple appearance models $zeta^k$ to compute the likelihood of $X_t$.
(4)We select the state $s_t$ which has maximum probability of $X_t$.
(5)Let $X_t$ represents the quantized HOG feature of the target at frame $t$, and then use Algorithm 1 to update parameters of multiple appearance models online in a streaming fashion.
```

6 Proposed Tracking Algorithm

Given the observation set of the target $X_{1:t} = [X_1, \ldots, X_t]$ up to time t, where each X_t represents a quantized HOG target feature at time t, the target state s_t(motion parameter set) can be determined by the maximum a posteriori(MAP) estimation as follows:

$$\hat{s_t} = argmax\ p(s_t|X_{1:t}) \tag{11}$$

where $p(s_t|X_{1:t})$ can be inferred by the Bayesian theorem in a recursive manner (with Markov assumption)

$$p(s_t|X_{1:t}) \propto p(X_t|s_t)p(s_t|X_{1:t-1}) \tag{12}$$

where $p(s_t|X_{1:t-1}) = \int p(s_t|s_{t-1})p(s_{t-1}|X_{1:t-1})ds_{t-1}$. The tracking process is governed by a dynamic model, i.e. $p(s_t|s_{t-1})$, and an observation model, i.e. $p(X_t|s_t)$.

A particle filter method [15] is adopted here to estimate the target state. In the particle filter, $p(s_t|X_{1:t})$ is approximated by a finite set of samples with important weights. Let $s_t = [l_x, l_y, \theta, s, \alpha, \phi]$, where $l_x, l_y, \theta, s, \alpha, \phi$ denote x, y translations, rotation angle, scale, aspect ratio, and skew respectively. We approximate the motion of a target between two consecutive frames with affine transformation. The state transition is formulated as $p(s_t|s_{t-1}) = N(s_t; s_{t-1}, \sum)$ where $\sum$ is the covariance matrix of six affine parameters. The observation model $p(X_t|s_t)$ denotes the likelihood of the observation X_t at state s_t. The Noisy-OR (NOR) [17] model is adopted for doing this:

$$p(X_t|s_t) = 1 - \prod_k (1 - p(X_t|s_t, \zeta^k)) \tag{13}$$

where $\zeta^k, k \in (1, 2, \ldots, K)$ represents the multiple appearance model parameters learned from Algorithm 1. The equation above has the desired property that if one of the appearance models has a high probability, the resulting probability will be high as well. Algorithm 2 illustrates the basic flow of our tracking algorithm.

Figure 2 shows how the online adaptive multiple appearances model working. These small face images show the appearance instance belong to each appearance model and the historical instances while tracking. The red rectangle in main frame is the tracking result based on our proposed model, and the green one is the ground truth. With the new tracking observation arrival, the number of clusters became varies.

7 Experiments

To evaluate our tracker, we compared the proposed tracker with 10 latest algorithms using 10 challenging public tracking datasets introduced by [20]. When evaluating the trackers, there are several problems should be discussed. We followed the evaluation methods from [20]. As object tracking is a traditional problem in computer vision, these trackers have quite different frameworks, so that all of them have advantage and disadvantage when meeting different challenges like occlusion and etc. Table 1 shows all the trackers (including our proposed algorithm) and their features and models. Note that in our proposed algorithm the HOG feature can be replaced by other features.

Table 1. Compare trackers and their representations in our experiment [20]

Trackers	Features	Models
LOT [22]	C	L
IVT [9]	PCA	H
ASLA [23]	SR	L, GM
L1ANG [25]	SR	H, GM
MTT [28]	SR	H, GM
VTD [24]	SPCA	H, GM
OAB [26]	Haar	H, DM
MIL [21]	Haar	H, DM
TLD [28]	BP	L, DM
Struck [27]	Haar	H, DM
AMAM [12]	Optional	DPMM
OAMAM	Optional	H, DPMM

One thing to emphasis is that all the trackers are running with adjusted parameters or simply use the parameters given by their publication for fair evaluation.

As mentioned before, a tracker might face tons of problems listed below in a real usage. According to the [20, 29], we divided these variation into six groups and analyzed some datasets by using this division. In the Table 2, we also add a short form of each challenge on each datasets. Here, the OCC stands for Occlusion, IV stands for Illumination Variation, R stands for Rotation which contains in-place rotation and out-of-place rotation, SV stands for Scale Variation while BC stands for Background Clutters.

One general problem for tracking is that the object may be occluded by other objects for several seconds. While in the dataset *Bolt*, the main object Bolt just kept the sportsman near him out in some of the frames and this will lead trackers to track on the sportsman near Bolt.

Table 2. Datasets and their problems

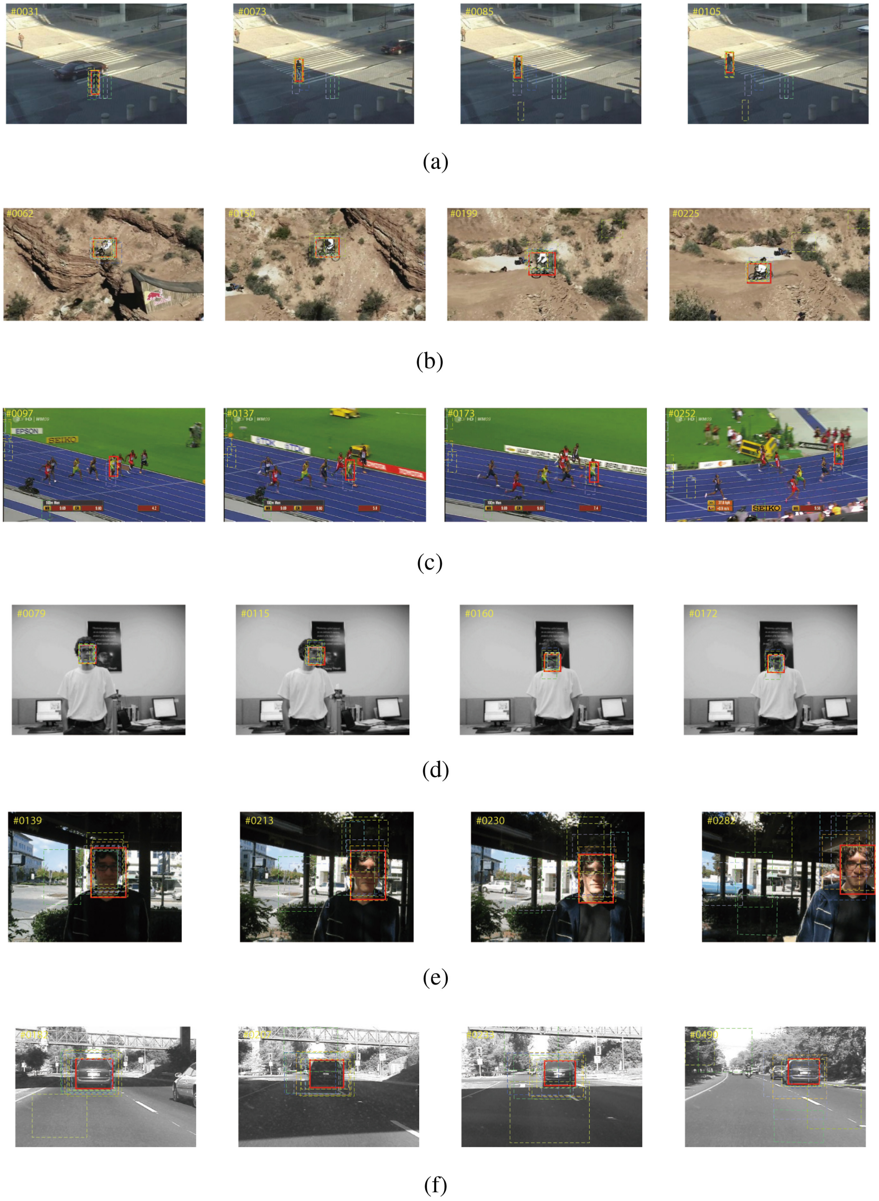
Dataset	Problems
CarDark	IV, BC
David2	R
Car4	IV, SV
Trellis	IV, SV, R, BC
Singer1	IV, SV, OCC, R, BC
Singer2	IV, R, BC
Bolt	OCC, R
Crossing	SV, R, BC
MountainBike	R, BC
Dog1	SV, R

This method we proposed didn't limited any certain kind of features for tracking task. Better features can get better tracking results. We simply applied HOG feature to implement.

It's common to use Center Location Error (CE [20]) and Overlap Score (OS [29]), to estimate the performance of the tracker. OS is calculated by the formula $score = \frac{area(ROI_T \cap ROI_G)}{area(ROI_T \cup ROI_G)}$. In the experiment, the $area(ROT_T)$ is the area of bounding box of tracking, and the $area(ROT_G)$ is the area of the ground truth. The CE is the Euclidean distance between the centers of tracking bounding box and the ground truth.

7.1 Online AMAM vs. Original AMAM [12]

In the previous sections, we compared our new proposed tracking method with AMAM tracking method [12]. As online method benefits the predicting speed on a long-term object, we compared these two methods in the time consumption. Figure 4(a) illustrates the DPMM time consumption of each frame in *Trellis* for both OAMAM method and AMAM method. It's obvious that AMAM method has a quite unaffordable time cost tracking for a long time while our online method performs relatively stable.



(a)

(b)

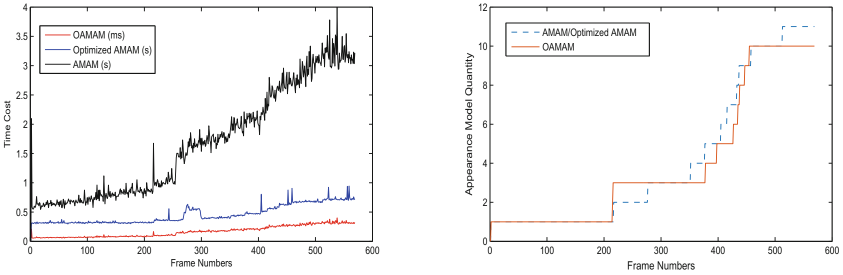
(c)

(d)

(e)

(f)

Fig. 3. All the images above are tracking results by trackers in Table 1 and dataset in Table 2, in which the bounding boxes in red are our results. (Color figure online)



(a) (b)

Fig. 4. (a) Is the DPMM time cost of each frame in *Trellis*, where OAMAM using *ms* and the other is *s*. (b) Is the quantity of appearance models of each frame when processing in *Trellis*.

Besides the time cost, they have a slightly difference in forming appearance models during the tracking task. Figure 4(b) shows the amount of appearance models for both methods in every frame in dataset *Trellis*.

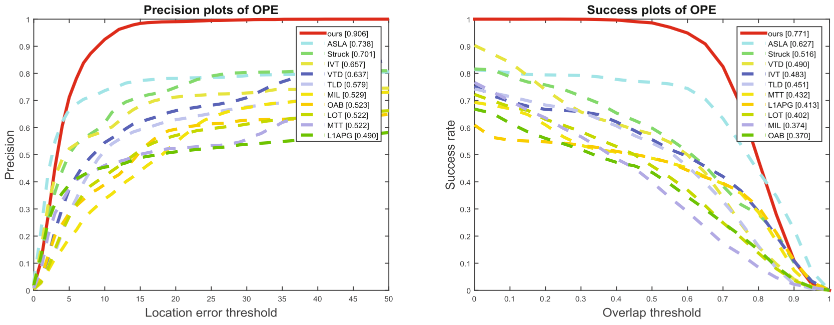
7.2 Qualitative Comparison

Our tracker has a robust performance while solving different challenges in different video sequences. Typical background problem can be seen in *MountainBike*, *Crossing*. In the Fig. 3 in *Crossing*, when a car was passing by the pedestrian, they shared similar dark colors in the frame 31 and result in the ASLA, Struck, and TLD's failure in tracking. In the frame 73 to frame 85 the target pedestrian blurred himself with the dark shade and only Struck, MIL and our tracker catched the target successfully(even Struck failed to track the pedestrian in the frame 31). In *MountainBike*, our tracker still performed well while the target was on the grass or dark shade in frame 62(VTD lost the target entirely from this frame), frame 150, frame 199, and frame 225. During the whole period of these two video sequences, our tracker tracked the target perfectly and constantly performed better than other trackers.

At the same time, there are view port varying problem in *Bolt* and rotation challenge in *David*2. In *Bolt*, the view port of the camera varied three times. It firstly lied in frame 97, as shown in the Fig. 3 Bolt was running towards the camera. The Second variation lied in frame 137 while Bolt was running parallel to the camera. The third variation lied in frame 252 while Bolt running away from the camera. Most of the trackers lost the target at the first stage, except four were still catching the Bolt. Only three trackers tracked Bolt successfully at the second stage. At the last stage, only our tracker was still working. In *David*2, there were abundant in-plane rotations and out-of-plane rotations. During the out-of-plane rotation(from the frame 79 to 115), half of the trackers had high CE rate even they did not lost the target.

In *Trellis* and *Car*4, there are significant illumination variations. In *Trellis*, The illumination of target varied from all dark to half dark during frame 139 to frame 213, and changed to bright in the frame 230. All the bounding box of these frames is shown in Fig. 3. In the frame 282 we could clearly find that only two trackers (ours and MIL) succeed in tracking the target while others drifted away because of the dark background. In *Car*4, the video sequences undergo serious illumination changes when the vehicle ran through a tunnel or under trees. At the frame 182, most of trackers performed well except two trackers fail to track the vehicle. But in the frame 207, 6 trackers enlarged its bounding box and drift away in frame 233 while the vehicle ran outside the tunnel. After the frame 490 and passed several trees and billboards, only 4 trakers including our tracker, MIL, ASLA and VTD were succeed in tracking target, and only our tracker didn't falsely enlarge its bounding box comparing to the ground truth.

We employed the protocol above to finish a comparison and analyzed all the data after evaluating. By adopting the OPE evaluation matrics, we compared the performance of trackers in all testing datasets with the same testing result shown in Fig. 5. From Fig. 5, we found that our tracking method outperforms state-of-art in the OPE evaluation on these 10 datasets. In the plot we can also infer that our tracking method is approximately 22 % better than the second best tracking method.



(a) (b)

Fig. 5. Success rate and precision of eleven trackers versus different thresholds under different attributions on ten video sequences.

8 Conclusion

In this paper, we proposed a new online adaptive multiple appearances model for long-term tracking. This approach remained more historical information on appearances of tracking target to avoid the target drifting or lost during the tracking caused by varying illumination or pose changing. We employed HOG to build the basic appearance representation of the tracking target in our algorithm framework. Multiple appearances representations were grouped unsupervised and dynamically by an online sequential approximation BNP learning method. Tracking result can be selected from candidate targets, which were predicted by trackers based on those appearance models, by using Noisy-OR method. Experiments on public datasets show that, our tracker has low variation (less than 0.002), low time cost for real-time tracking, and high tracking performance (22 % better than other 10 trackers in average) when compared to the state-of-the-art.

References

1. Li, Y., Ai, H., Yamashita, T., Lao, S., Kawade, M.: Tracking in low frame rate video: a cascade particle filter with discriminative observers of different life spans. IEEE Trans. Pattern Anal. Mach. Intell. **30**(10), 1728–1740 (2008)
2. An, S., An, D.: Stochastic relaxation, Gibbs distributions, the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
3. Blei, D.M., Jordan, M.I.: Variational inference for dirichlet process mixtures. Bayesian Anal. **1**(121–144), 1 (2005)
4. Stauffer, C., Grimson, W.E.L., Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 246–252. IEEE, Los Alamitos, August 1999
5. Lin, D.: Online Learning of nonparametric mixture models via sequential variational approximation. In: Proceedings of NIPS (2013)
6. Ulker, Y., Gunsel, B., Cemgil, A.T.: Sequential Monte Carlo samplers for Dirichlet process mixtures. In: AISTATS 2010 (2010)
7. Cannons, K.: A review of visual tracking. Department of Computer Science, York Univ., Toronto, ON, Canada, Technical report CSE-2008–07 (2008)
8. Jepson, A.D., Fleet, D.J., El-Maraghi, T.F.: Robust online appearance models for visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1296–1311 (2003)
9. Ross, D., Lim, J., Lin, R.-S., Yang, M.-H.: Incremental learning for robust visual tracking. IJCV **77**(1–3), 125–141 (2008)
10. Avidan, S.: Ensemble tracking. PAMI **29**(2), 261–271 (2007)
11. Babenko, B., Yang, M.-H., Belongie, S.: Visual tracking with online multiple instance learning. In: CVPR, pp. 983–990 (2009)
12. Tang, S., Zhang, L., Chi, J., Wang, Z., Ding, G.: Adaptive multiple appearances model framework for long-term robust tracking. In: Ho, Y.-S., Sang, J., Ro, Y.M., Kim, J., Wu, F. (eds.) PCM 2015. LNCS, vol. 9314, pp. 160–170. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24075-6_16
13. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of ICCV, pp. 1150–1157 (1999)
14. Ferguson, T.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**(2), 209–230 (1973)
15. Isard, M., Blake, A.: CONDENSATION-conditional density propagation for visual tracking. Int. J. Comput. Vis. **29**(1), 5–28 (1998)
16. Pitman, J.: Combinatorial Stochastic Processes. Lecture Notes in Mathematics, vol. 1875. Springer, Heidelberg (2006)
17. Viola, P., Platt, J.C., Zhang, C.: Multiple instance boosting for object detection. In: Proceeding of Neural Information Processing Systems, pp. 1417–1426 (2005)
18. Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley, New York (1994)
19. Bissacco, A., Yang, M., Soatto, S.: Detecting humans via their pose. Adv. Neural Inf. Process. Syst. **19**, 169 (2007)

20. Wu, Y., Lim, J., Yang, M.: Online Object Tracking: a benchmark. In: Proceeding IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2411–2418 (2013)
21. Babenko, B., Yang, M., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **33**(8), 1619–1632 (2011)
22. Oron, S., Bar-Hillel, A., Levi, D., Avidan, S.: Locally orderless tracking. In: CVPR (2012)
23. Jia, X., Lu, H., Yang, M.: Visual tracking via adaptive structural local sparse appearance model. In: CVPR (2012)
24. Kwon, J., Lee, K.M.: Visual tracking decomposition. In: CVPR (2010)
25. Bao, C., Wu, Y., Ling, H., Ji, H.: Real time robust L1 tracker using accelerated proximal gradient approach. In: CVPR (2012)
26. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via online boosting. In: BMVC (2006)
27. Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M.M., Hicks, S., Torr, P.: Struck: structured output tracking with kernels. In: ICCV (2011)
28. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Robust visual tracking via multi-task sparse learning. In: CVPR (2012)
29. Everingham, M., Van Gool, L., Williams, C.K.I., et al.: The Pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)